DOCUMENT RESUME

ED 207 172 EA 013 914

AUTHOR Becker, Wesley C; And Others
TITLE Design and Measurement Issues in Follow Through Research.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 79
CONTRACT RF78-101
NOTE 37p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; Analysis of Covariance; Control Groups; Disadvantaged Youth; Educational Research; Elementary Education; Federal Aid; Followup Studies; Measurement Techniques; Models; Outcomes of Education; Preschool Education; *Program Effectiveness; *Program Implementation; *Research Design; *Research Needs; *Research Problems; Statistical Analysis
IDENTIFIERS *Project Follow Through

ABSTRACT

The focus of future Follow Through research should be oriented toward showing what can be done in order to offer information to both federal and local planners on which systems of curriculum design, classroom management, and parent involvement appear to be most effective. To assure this goal, the issues of program efficacy and implementation should be experimentally isolated in the next phase of Follow Through research. Separate studies should be conducted to explore effective implementation strategies in big cities. The new Follow Through design would include improvements over the former design: more care in selecting local comparison groups; better assessment of entry performance of children; and assessment of implementation processes in Follow Through and control classrooms. Within this design, separate component analysis studies are needed. Appendix I details potential component analysis studies. There is also a need for Follow Through to be extended into grades 4 through 6, and for selected sponsors to demonstrate solutions to issues in the intermediate grades. The best approach towards broad range assessment would be to administer a yearly test battery to all full-term Follow Through students and to conduct small scale longitudinal studies. (Author/MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED207172

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Design and Measurement Issues in Follow Through Research[1]

Wesley C. Becker

Russell M. Gersten

Douglas W. Carnine

University of Oregon

This research was supported by U.S. Office of Education Contract # RF78-101. The views expressed herein do not reflect the views of the funding agency.

1979

EA 013 914

Future Research Studies at OE

Goals

The Office of Education should be conducting research which will aid the Congress and Educational Planners to make better decisions in the utiliation of Federal resources to improve education. OE should also be conducting demonstration studies whose outcomes will assist school districts all over the country to make better decisions about classroom programs, management procedures, parental involvement, finance, etc. A major goal of the kinds of applied research OE should do is showing what can be done to improve problem areas where school currently fail the most. The demonstrations must be strong enough to serve as a catalyst or prod for others to do what is necessary to produce equally effective outcomes for children.

1. Major demonstrations are needed:
   a. with disadvantaged children similar to those now in Follow Through, with better control and selection of comparison groups; better assessment of entry performance measures; and assessment of implementation processes.
   b. extending Follow Through to grades 4, 5, and 6 for selected sponsors to demonstrate solutions to current intermediate level problems.
   c. using the planned-variations strategy in bilingual education,
   d. large-scaled demonstrations in one or more Big Cities (using Title I funds). Models must have been demonstrated to be effective in controlled settings. Sponsors must present realistic strategies for Big City implementation.

2. Component analysis studies across models (or within models) are needed to:

   a. show the effects of educational programs independently of management systems,

   b. show the effects of various components of management systems in increasing learning opportunities for children,

   c. show the effects of various kinds of procedures to involve parents as teachers of their children,

   d. show the independent effects of different subject area components on student outcomes.

   e. show the effects of different levels of funding on program outcomes.

(Appendix 1 details potential components analysis studies.)

In the major part of this discussion we will deal with the issues identified by OBPE in the RFP 78-101 as they apply to possible new demonstrations of a Follow Through Planned-Variation type. With that background, it will then be possible to discuss the other kinds of research problems of importance more briefly.

1.a. A New Follow Through Designs

The evaluation results would be a lot clearer if a true experimental design (with random assignment of kids to treatment etc.) was utilized, rather than a quasi-experimental design. There is much less faith in analysis of covariance adjustments (Campbell & Boruch, 1975; Campbell & Erlbacher, 1970; Elashoff, 1969) or multiple covariance adjustments, in 1979 than in 1970-71 when the Follow Through evaluation was designed. However, for political, social, and ethical reasons, random assignment does not seem feasible. An alternative is the regression discontinuity model

(Campbell & Stanley, 1966; Horst Tallmadge & Wood, 1975). While possibly feasible for Title I evaluations, it is not appropriate for Follow Through entry criteria. Regression discontinuity avoids the Campbell-Erlbacher regression artifacts problem, but requires a clear cut-off score on some test(s) to determine admission to the program. FT guidelines for admissions just do not currently fit. In addition, there are several other problems with regression discontinuity: (1) assumption of a linear pre-post test relationship for all kids and all SES levels (which is likely untrue for MAT reading), and (2) the model has not been widely used and there may be unknown problems.

A quasi-experimental design with use of multiple covariates to adjust outcomes with initial differences between groups can be improved (over the SRI--Abt evaluation) by:

1. Insuring fair, comparable local control (or contrast) groups.
2. Insuring that each sponsor has roughly equivalent ethnic-regional groups, so that fair contrasts between sponsors can be made.
3. Improving assessment of entry level skills.
4. Taking process measures for both experimental and control groups.
5. Eliminating districts with potentially high implementation problems.

Hopefully, any covariance adjustments that need to be made will be minimal. Related alternative designs are discussed under Comparison Groups.

Unit of Analysis

There are four contenders: (a) the child, (b) the classroom, (c) the school, and (d) the site. Two of these have proved infeasible.

The classroom is an infeasible choice of unit of analysis for a longitudinal study because (a) the use of ability groupings in many schools,

and (b) changes in teachers and students from year-to-year. The former contributes to huge "error variance", and consequently low statistical power; the latter makes the classroom inappropriate for a longitudinal (as opposed to a cross-sectional) study. The school was utilized by Abt with little success. Again, SES variability between schools creates large error variance, and low power. In addition, variability in the number of schools per site (which is purely a function of community size and policy) can lead to unequal weighing of certain sites.

Thus, we are left with the child and/or the site as the unit. We will argue that the data be analyzed both ways, but that the following issues be considered:

The site is favored by House et al., (1978)and Bereiter & Kurland (1978). It fits the assumption of independence of experimental units necessary for statistical tests (Peckham, Glass & Hopkins, 1969). Furthermore, attrition does not as critically affect the power analysis since only mean pre-test and post-test scores for a site will be used.

There are two major problems, though--one statistical, one theoretical. As Haney (1974, 1977) points out, using the site as a unit drastically reduces the N, so that an experiment with perhaps 50,000 kids may have problems with extremely low power due to small sample size. (However, this problem is usually compensated by smaller error terms.) Low power would be especially likely if there are strong regional differences, creating a situation of small N and large error variance leading to weak power for statistical significance. However, blocking by ethnic-regional groups can take out ethnic-regional variation. Peckham et al. (1969) claim that a sample size of 6 or more often gives reasonable power for classrooms as units. but there is a need to ascertain whether the same is true when widely varying sites are used. The Bereiter & Kurland solution adds power to the analysis by including all sponsors in the same analysis.

The second issue Haney (1977) raises (rather eloquently) is that Follow Through "treats" children, not schools, not town or cities. The goal of FT is to have impract on disadvantaged children; any effects on the community are secondary. Whereas Gall (1977), Peckham, et al. (1969) argue that in classroom research the instructional group, not the child, is the experimental unit (because kids affect each other during classroom instruction), the case in Follow Through is not exactly analogous. In four years, each child interacts with a complex web of teachers, aides, changing classmates, guidance counselors, etc. There is no clear system or pattern to these interactions. Thus, we could argue that two kids who are in grade 3 in one site have approximately the same relationship as 2 college sophomores who sign up for a laboratory psychology experiment - an unknown pattern of dependence and independence.

The major drawbacks House et al.(1978) find in using the child as a unit are:

1. The units of observation are not indeperdent, violating statistical assumptions.
2. The larger the site, the more likely the chance of finding statistical significance and bias the findings in favor of sponsors in larger sites.

The House et al. argument has one serious flaw--"hypersensitive" statistical tests, with large Ns, are equally likely to declare (educationally) non-significant effects statistically significant in either direction, i.e., either favoring FT or favoring the contrast group. Nonetheless, there is a need to: (1) omit very small and very large sites; (2) set an optimal N per site (eg., N = 100 per grade per year; (3) report educational significance and ω^2 for each site. We recommend the use of site, blocked by sponsor and ethnic-regional groups, for statistical

analysis, and the use of the child for norm referenced gains and other longitudinal analysis.

Comparison Groups

Perhaps the most serious problem in the Abt-SRI Follow Through evaluation was poorly chosen, inappropriate "control" groups. The control groups were nominated by local district staff and turned out to have largely less disadvantaged children. In addition, some "control" groups were involved in innovative programs under Title I or III. The lack of any process measures, or fidelity of treatment measures on "control" groups, led to more obscurity.

If comparison groups are to be used, we recommend the following:

1. Careful selection of control groups by an outside panel so that they are: (a) from similar SES background and (b) not involved in special experimental approaches except by design.
2. The use of other sponsors as contrast groups in some sites along with "control" groups (i.e., several sponsors represented in one community
3. The use of Title I ethnic-regional pooled comparison groups, and if appropriate, use of Follow Through control groups, pooled for ethnic-regional groupings.
4. The use of publisher norms (especially if ethnic-regional norms could be obtained from their standardization data pools) to provide norm-referenced comparisons of outcomes. This has become especially important, in that recent versions of standardized tests (the 1975 CTBS, 1977 CAT) use a much larger proportion of minority children in standardization samples. The proportions vary from publisher to publishers; but this information is almost always explicit in the technical bulletins.

(Other control groups issues are discussed in the next section.)

Premeasures and Covariates

Better and more timely measure of student entry skills is needed. Furthermore, when the children are in special preschool programs, ideally the premeasures should antedate those programs. Where premeasures can be obtained, they can be used to make best estimates of post preschool levels.

A pre-battery should include: (a) the Metro Pre-Primer or the CTBS or the CAT Pre-Primer, and (b) the Boehm (1971) test of concepts and/or the Assessment of Children's Language Comprehension (Foster et al, 1973). Additional data for possible use as covariates should be gathered on parent income, parent education, whether English is the child's primary language, ethnic group, and nature of preschool experiences. Where Spanish is the primary language, the children should be pretested in both English and Spanish on fair forms of the above tests. The premeasures and background measures may be used as covariates.

The use of analysis of covariance was one of the most controversial aspects of the Abt analyses. Since a true experimental design is infeasible, there will likely be a need for some covariance adjustments. However, the proposed design attempts to drastically reduce the magnitude of the adjustments by choosing contrast groups as similar as possible to Follow Through groups, and blocking by sponsor and ethnic-regional groupings. Bereiter & Kurland's (1978) suggestion of using the contrast group posttest score(s) as covariates seems a sound one.

The ultimate choice of covariates must be an empirical decisions, not an arbitrary decision. We see two possible options. In the first, a correlation matrix between potential covariates and posttest scores is computed, and the three covariates which explain the largest (non-redundant) proportion of the variance are selected.

There would definitely be a need for cross validation procedures. Ideally, these would be done on an independent sample of similar children. If that procedure is infeasible, the sample should be randomly split into subsamples for purposes of cross-validation. The technique of covariance structure analysis (Kerlinger, 1979) could also be explored. If properly conducted, the correlations between entry behaviors, home environment variables, type of treatment, and student outcome measures will supply important information about the complex relationships among these factors.

Threats to Validity

1. Between-sponsor comparisons. The problems of large regional ethnic differences in entry skills and family characteristics of children argues for classifying possible participating communities in to ethnic regional groups, and then blocking by regional groups for national comparisons, or studying sponsor effects within regions. We suggest the following major groupings:

Chicago-Southwest

Rural Blacks--Southeast and Mid-South

Urban Blacks--Northeast and Midwest

Working Class Whites--Midwest and Farwest

Native Americans-- ??

We also recommend a primary focus on middle-size communities (50,000 to 750,000), expect for limited implementation studies for large cities.

2. Acts of Unions, governments and ???. In each project site, a contract should be written with a secretary (or equivalent) to log daily the factors which actually disrupt the possibility of teaching going on, and the potential quality of the teaching:

1. Integration procedures which limit "pull out" time, or whatever.
2. The closing of schools. The busing to new schools, etc.
3. Strikes.
4. Weather, and other non-teaching days.
5. Teacher staff reductions, resignations, and other factors requiring retraining.
6. Replacements of key staff.

As much as 20 to 30 percent of a program's effects could be attenuated by such happenings. They could be used to extrapolate program effects against a constant 175 day teaching year, or whatever.

Degree of Implementation

Both measures of quantity of instruction delivered (i.e. academic engaged time) to the children and quality of instruction are needed, for both experimental and comparison groups. To adequately assess these variables, we propose a three-tiered model.

The first tier would consist of fairly crude (and inexpensive) checklists, which an independent evaluator could fill out after a brief observation of the classroom and a structured interview with the teacher (and perhaps a teacher supervisor). These checklists would deal with the very basic components of a model. For example, a checklist for the Direct Instruction model would include items dealing with type of reading material used (phonics-based versus basal readers), whether small group instruction is used, whether a mastery learning paradigm with frequent testing of students' oral responses is used, etc. A Parent Education model checklist would focus on frequency of home visits. A Behavior Analysis checklist would focus on frequency of data collection, whether teachers systematically use reinforcement and token economics, etc. These checklists would be a crude way to distinguish between experimental and

comparison groups to insure that comparison groups are different in important ways from Follow Through classrooms. Furthermore, they would provide a crude measure of whether Follow Through classrooms are in fact implementing the major components of the model.

The second tier would also be a combination checklist-observation form, but a refined, carefully developed and validated one. For this form, observers might visit a classroom three times before an assessment is made. Much more specific information would be obtained - such as average number of minutes spent on reading per day, rate and quality of teacher's corrective feedback to students, amount of time spent in small group instruction vs. individual seatwork. Also, adequacy of supervision and training, a degree of administrative support for the model, parent involvement and contribution to child learning would be assessed. This should be done for all experimental and comparison classrooms at regular intervals (four times a year). The items and/or variables observed should be chosen with care. We would recommend some concurrent validity studies (using supervisor ratings) before the final format is chosen. The focus would be on variables sponsors feel are crucial to the approach, but which can be operationalized in some form, such as academic engaged time for Oregon, "desirable teaching behaviors" for Chapel Hill. We see this instrument as being much less costly to administer (and, in some ways, more focused) than a Stallings-type observational instrument. Thus, every Follow Through and comparison classroom could be observed and assessed, perhaps four times a year.

Many variables should be constant across sponsors--eg., amount of time per day spent on reading, type of math curriculum used, quality and degree of home involvement. Others would be model-specific. A checklist

format would seem a valid way to secure important information such as type of readers used, frequency of field trips--information which could not emanate from an observational code.

The third tier would involve frequent (i.e., perhaps 20 times per year) direct observation of children in selected classrooms. This procedure would be reserved for specific small scale studies (see Appendix I). Two possible avenues are open, and--depending on the nature of the research questions--we believe both should be pursued. The first is the aforementioned Stallings (1973, 1977) observational system. The major advantage of this type of code is that it can be used across all models, experimental and control group. The major disadvantage is that some of the nuances of each methodology--eg., sequencing of examples in DISTAR, use of correction procedures--are likely to be lost.

An alternative is the development of sponsor-specific codes. The Direct Instruction Supervisory Code (Paine & Gersten, 1979), developed to access implementation of the Direct Instruction model in San Diego, is an example. Five critical teaching behaviors in Direct Instruction methodology (pacing, use of correction, use of mastery learning, accuracy of formats, accuracy of signals) were operationalized, and a behavioral scoring code developed. Teachers are observed for 12 minutes, and frequencies of each of these behaviors is recorded. A mini-generalizability study is now being conducted to determine the minimum number of observations necessary to obtain a reliable measure of teacher performance. Concurrent validity is also being assessed by correlations with supervisor ratings. Results appear to be promising (though refinement is necessary). Similar efforts have been attempted by Georgia State and Chapel Hill (among others), and we see refinement and development of these performance measures as a crucial component in small scale research.

Summary

The focus of future Follow Through research should be oriented toward showing what can be done, to offer information to both federal and local planners on which systems of curriculum design, classroom management, and parent involvement appear to be most effective. The demonstrations must be powerful enough to serve as catalysts for education change. To assure this goal, the issues of program efficacy and implementation should be experimentally isolated in the next phase of Follow Through research. Demonstrations of program efficacy should be limited to communities where there is a reasonable chance of effective implementation. This would probably involve large towns and small or moderate-sized communities (50,000 to 750,000). Separate studies should be conducted to explore effective implementation strategies in big cities, using models that have previously demonstrated effectiveness. The new Follow Through design would improve on the former design in several respects: a) more care in selecting local comparison groups, b) better assessment of entry performance of children, and c) assessment of implementation processes in Follow Through and "control" classrooms. In addition, Title I achievement score data (both local and regional) would be used to supplement comparison group data. The data analysis design of Bereiter and Kurland (1978) offers a sophisticated model for site level analysis. They used analysis of covariance to contrast performance of all sponsors, using the site as a unit of analysis and comparison group scores as covariates. We believe this approach, in combination with the use of norm-referenced comparisons, will offer the best first cut at evidence of model effectiveness. To effectively deal with the issue of inter-region variability, we urge that all participating communities be classified into one of the following major groupings:

Chicanos - Southwest
Rural Blacks - Southeast and Mid-South
Urban Blacks - Northeast and Midwest
Working Class Whites - Midwest and Farwest
Native Americans

Each sponsor should be equally represented in each "region." This would allow for the statistical procedure of blocking to reduce unexplained variability and for an analysis of interactions between models and type of children (eg., whether a specific model is particularly effective with Urban Blacks). By distributing sponsors equally across regions, the need for multiple covariance adjustment should be minimized, as each sponsor will have roughly equivalent samples. It would be perfectly appropriate, and advisable, to include sites of more than one sponsor in a given community.

Within this grand design, separate component analysis studies are needed to investigate issues such as: (a) the effects of educational programs independently of management systems, (b) the efficacy of various parent involvement procedures, (c) the effects of various management systems on learning opportunities of children. Separate small scale studies should examine: (a) the effects of various model or bilingual populations and (b) efficacy of Follow Through with handicapped and migrant populations. There is also a need for Follow Through to be extended into Grades 4 through 6, and for selected sponsors to demonstrate solutions to issues in the intermediate grades.

Outcome Measures

We believe the best approach towards a broad range assessment would be to: (a) administer a yearly test battery to all full-term Follow Through students using standardized, psychometrically validated instruments in academic content areas and (b) evaluate longitudinal changes in order to measure areas such as oral language, self concept, and written expression in small scale studies using perhaps 100 children per sponsor per year, with an equal number of local control group children. (Possible savings - omit post 2nd testing on standardized achievement tests.)

A. Achievement Test Battery

The achievement test battery should be a standardized test battery using a reasonably representative standardization sample, with suitable reliability indexes. The test should be administered shortly after entry in Kindergarten (eg., within the first 3 weeks), and at the end of each academic year. In this way, norm-referenced comparisons (Horst, Tallmadge & Wood, 1975) as well as control group comparisons can be made. Tests should be administered at the appropriate standardization dates, except for pre-Kindergarten. It is crucial that pre-K tests assess entry skills, rather than be confounded with treatment effects. Therefore, we would urge pre-K tests be administered in the second week of school, and individually, even though this differs from standardization test dates. (As long as the same policy is used for all sites, there will be no bias.)

Important concerns in the choice of test batteries is content analysis. A secondary, but quite serious concern, is the suitability of test format and instructions for disadvantaged children. The content analyses should be done before the evaluation is designed, not afterwards. (Appendix 2

is an example of a first try at the type of content analysis that sponsors should perform prior to evaluation.) To assure uniformity of familiarity with test formats, we urge that Follow Through design a "practice kit", which all kids--experimental and control--use each year. No other practice with test formats will be permitted. (We will be glad to assist in the design of such a kit.)

There appear to be four main contenders for a primary test battery - the Metropolitan Achievement Test, the California Achievement Test, the CTBS, and the Stanford Achievement Test. Current versions of these tests have a pre-K form (except for the Stanford). Our current, tentative conclusion is that the Metropolitan-including the Reading Instructional Tests (Farr et al., 1978)-would be the superior choice because of: a) general clarity of instructions and format, and b) the pre-K test appears to be the most educationally relevant (eg., less focus on mazes and the like), and c) the third grade reading subtests cover a wider range of skills and processes than any of the othres, including vocabulary, grapheme/phoneme associations, words in context, and reading comprehension.

1. Problems with the Metropolitan

The Math Problem Solving and Math Computation subtests seem reasonable measures of most math programs; however, Math Concepts is considered by Bereiter & Kurland (1978) to be "a hodge podge of new math and old math." The Buros reviewer (Wolf, 1978) considers this subtest equally weak, containing several items completely unrelated to mathematics. Taken as a whole, however, coverage of math is more than adequate. Similarly, the Language and Spelling subtests seem to assess what they purport to do reasonably well.

The main problem is with Reading, particularly on the Elementary form of the test, given at the end of grade 3. Becker (1978) has pointed out

that the 3rd grade form of the MAT (as well as the CAT, CTBS, and SAT), unlike first and second grade forms of the tests, uses virtually an uncontrolled (average adult) vocabulary. This seems to put low-income kids at a disadvantage, since schools leave the primary teaching of vocabulary concepts to the home environment, and many low income kids do not get the type of language stimulation that middle class kids do. We therefore think MAT Reading (Elementary) should be supplemented by: (a) a reading comprehension test using a controlled third grade vocabulary, (c) an actual test of children's vocabulary. These would provide more detailed information on relative strengths and weaknesses of reading and language programs (details are presented later). To summarize:

a) A standard achievement test battery should be used.

b) A content analysis of tests should be done before the evaluation is conducted (rather than after, or not at all) to ascertain degree of fit between tests and program goals.

c) Measurement of entering skills in kindergarten was a weak point in the past evaluation. The WRAT is too narrow. A preprimer form of a standardized test battery (such as MAT, CTBS) given individually, would provide a broader band assessment of entry skills. These should be supplemented with a test of a child's basic knowledge of simple concepts (eg., Boehm, Assessment of Children's Language Competencies, Basic Concept Inventory), since

readiness tests omit this area, and it's an important index of prior learning.

d) There should be uniform practice in test-taking across all sites, experimental and control.

e) There is a need to supplement 3rd grade forms of standardized achievement tests with other measures of reading--oral reading rate and fluency, reading comprehension with controlled vocabulary.

In addition:

f) Raven's Progressive Matrices should be dropped from an achievement test battery. It does not measure the goals most communities have for Follow Through. The Raven's skills can be taught, but they are not as important as related language reasoning skills.

g) Self-concept tests should not be part of a national evaluation battery. Reliability and validity are weak (eg., House et al, 1978); and there is little evidence they measure what they purport to do. More intensive research on small groups of children should be conducted to look at changes in self concept.

2. Additions to Achievement Test Battery

a. Test of Basic Concepts

A test of understanding of basic concepts - such as "in," "under," "around" - would be a useful, cost-efficient addition to the kindergarten assessment battery. Either the Boehm Test of Basic concepts, ACLC (Assessment of Children's Language Comprehension) or the Test of Language Development would measure what kids have learned at home. The content of these tests fits program goals for Follow Through that most communities would have for disadvantaged kids. This area is not adequately covered in preprimer versions of standardized achievement tests. Probably only one posttest on this measure

(at the end of kindergarten) need be given, if any. Data from these tests would be more useful as pretest measures (i.e., covariates) than as outcome measures. (The ACLC and TOLD is being used on a pilot basis in San Diego in 1979.)

b. Additional Reading Measures

The battery should include a measure of oral reading. The main rationale is that there is no evidence that standardized tests such as the MAT adequately measure decoding or oral reading. Probably the most economical procedure would be to obtain a stratified random sample of 200-300 students per sponsor per cohort. Responses would be taped and scored by impartial judges. Judges should be trained to high levels of reliability. The test would be given to end of grade 3 students only.

There are 3 main contenders - Grey, Gilmore, Spache. The Spache seems the best since: (a) it is not as heavily rooted in meaning-based basals as the others, (b) the norms are the most recent (1974). The main drawbacks to the Spache are: (a) small N for standardization (b) quality of the comprehension questions. Another alternative - and one we would strongly endorse - is to develop a new measure of oral reading.

There also is a strong need to develop a measure of reading comprehension using a controlled vocabulary. This would allow for a much more accurate evaluation of areas of strengths and weaknesses of various programs. Finally, a specific measure of third grader's word knowledge (as opposed to word knowledge confounded with reading ability, as in the MAT) would be a useful addition. The work of Dupuy (cited in Becker, 1978) offers a useful starting point and provides possible test items.

c. Small Scale Studies

Certain educational goals--such as oral language, self-concept, written expression-- are important, yet difficult (and expensive) to assess. We feel the best route in these areas is to intensively examine small groups of FT children (with local controls) in a longitudinal design.

1. Self-Concept. Primarily, because no one knows what it is, let alone how to measure it, self-concept has been the biggest headache in instrumentation. The original Abt evaluation contained 3 affective measures: (a) the Coopersmith (1967), a measure of self-concept, and (b) 2 measures of locus of control, IARS (+) and IARS (-). All these instruments are controversial. There is a general feeling that some measure of self-concept and/or social behavior be included. However, the technical information (temporal stability, validity) on self-concept measures is poor. They are subject to examiner influence and limited by the child's receptive language; the test format is quite contrived. House, et al. (1978) suggested they be thrown out of the test battery.

There are 5 other potential tests, although it's unclear if or how they would be any better than the Coopersmith: (a) the Tennessee, (b) the Piers-Harris, (c) the Brown IDS self-concept (uses polaroid picture of kid) (d) the Boles Pictorial Self Concept (based on Q-sort methodology), and (e) the "About Me" (cited in Borich). However, there is no evidence or logic that suggests they will be superior to those used by SRI.

We recommend that no self-esteem test be part of a test battery. Instead, self-esteem and social behavior should be measured by several small-scale experiments. Perhaps the same kids could be used for both a self-esteem and an oral language study. The focus of the study would be on naturalistic observations of kids in schools during a structured time (probably a group instruction setting) and during a free play or recess time. We recommend selecting randomly 8 kids from 10 classrooms and observe them four times a year in both settings. Actual verbal behavior and a coding of other behaviors would be recorded. The work of Hill Walker on peer interaction might serve as a methodoligical point of departure (Hops et al., 1978). Procedures from the Behavioral Analysis Follow Through Model should also be examined as well as the work of de Charmes (1976).

Sampling would try to include some Blacks, some Chicanos, some Anglos, and some Native American kids. Each "experimental" sample would have a local "control" sample to account for regional variation in behaviors patterns. Obviously teachers have a powerful effect on child behavior; we should be getting a reasonable range of teachers (10 per year for 4 years). This study would offer fascinating longitudinal data. There are a multitude of problems, eg., how to analyze verbal statements, how to analyze interaction patterns etc., etc. We may want to alsc include some "clinical" interview data, perhaps even simultaneously. We would strongly recommend a two-year pilot study now to come to grips with measurement problems. (See Appendix for description of our pilot study.)

It would enhance the study if the aforementioned set of self-concept measures were administered, along with teacher and peer ratings. This would provide a set of concurrent validity measures that could offer valuable information for the field self-concept measurement in general.

2. Language Skills. Again, a small N longitudinal study is recommended. There seem to be problems with all existent instruments. There is a sentence repetition test developed by High Scope (Anastasiow & Hanes, 1974) with low reliability. However, an oral language instrument should be part of the study. Standard receptive language tests (such as the Carrow, or the Peabody Picture Vocabulary Test) may be included, but they only tap a small portion of what is oral language. The Test of Language Development (Newcomer & Hammill, 1977) appears to be an improvement. It taps the widest range of language activities, and has reasonable technical characteristics. It is being piloted in the San Diego site in May, 1979. However, only some of the subtests appear to be relevant to the program goals of most sponsors. We would probably recommend that only

the Oral Vocabulary and Sentence Imitation subtests we included in an evaluation battery. The other subtests seem too heavily loaded with middle class, almost quaint sentence constructions.

Language assessment is one area where instrument development is a priority. There is a need for an exhaustive review of current language assessment instruments, including those that look at English as a Second Language. Our sense is that most language assessment is clinically oriented, and there is little broadly relevant technical information available (reliability, validity, and even scoring). A possible base for instrument development is this:

a) Elicited Speech

1. Sentence repetition.
2. Responses to a set of specific questions.
3. Picture descriptions.
4. Something that is comparable to the Boehm Test of Concepts, except expressive (productive) rather than receptive. This would deal with concepts of causality, description of multiple attributes, logical deductions, and a brief narrative based on a small set of prompts.

b) Spontaneous (emitted) speech

How to assess this presents an extreme problem. There are techniques for language sample analysis according to: (a) phonological, (b) syntactic, (c) semantic, and (d) pragmatic construction. Little of this work is easily quantifiable, let alone psychometrically developed.

Coordinating the oral language and self-concept studies might make sense. It could be cost efficient, and also offer some information on the relationship between language skills and self-concept.

c) Writing Skills

There is also a need to consider allocating funds for the development of an instrument to assess writing skills. Currently available standardized tests of writing appear restricted in scope to proofreading words and phrases. On the other hand, instruments designed to analyze creative compositions (for example, the effort of High Scope) are usually costly to administer and difficult to score reliably. Any test of writing skills which might be developed should require production of written material as well as proofreading printed matter. Further, the instrument should assess such skills as appropriateness and fluency of ideas, as well as mechanical skills, such as punctuation, capitalization, and grammatical accuracy. Care should be taken to design an instrument which can be easily administered and scored.

Summary

The best approach towards broad range assessment would be to:

A. Administer a yearly test battery to all full-term Follow Through students using standardized, psychometrically validated instruments that meet community and Federal educational objectives. Thorough content analyses should be conducted on all instruments by all sponsors prior to the evaluation, rather than afterwards. An achievement test battery such as the Metropolitan Achievement Test or CTBS or California Achievement Test have several positive aspects: reasonably representative standardization samples so norm-referenced comparisons can be made, adequate reliability, and reasonable coverage of the areas of math, spelling, and language mechanics. However, there is need to supplement thess tests in several ways. First, entry assessment should be supplemented by a measure of the child's under-

standing of basic language concepts in both English and Spanish. Several tests are currently available. Second, third-grade reading needs a broader band assessment than any of the achievement tests offer. They should be supplemented by: (a) an oral reading test to assess decoding ability and fluency, (b) a test of reading comprehension using a controlled vocabulary, and perhaps (c) a test of the child's basic word knowledge.

B. Conduct small scale longitudinal studies to focus on self-concept, oral language and, to a lesser extent, written expression. Since there are serious weaknesses in all extant instruments in the field, Follow Through must develop its own. The best route would seem to be on naturalisitc observation of children in schools during both a structured time and a loosely structured time. Interaction rates, observations of play behavior, and analysis of the child's actual verbal statements would comprise the key components of these instruments. In the area of oral language development, some structured and semi-structured testing situations (eg., a sentence repetition test, tests involving the child's ability to describe a series of pictures, etc.) should also be included.

References

Anastasiow, N.J. & Hanes, M.L. Sentence repetition task. Unpublished manuscript. Institute for Child Study, Indiana University, 1974.

Becker, W. C. Teaching reading and language to the disadvantaged--What we have learned from field research. Harvard Educational Review, 1977, 47, 518-543.

Becker, W. C. & Engelmann, S. Analysis of achievement data on six cohorts of low income children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model. Follow Through Project, Technical Report #78-1. Unpublished manuscript, University of Oregon, Eugene, OR, 1978.

Bereiter, C. & Kurland, M. A constructive look at Follow Through results. Unpublished manuscript. Toronto: Ontario Institute for Studies in Education, 1978.

Bock, G., Stebbins, L. B. & Proper, E. C. Education as experimentation: A planned variation model (Vol. IV,A). Cambridge, Mass.: Abt Associates, 1977.

Boehm, Ann E. Boehm Test of basic concepts. New York: Psychological Corporation, 1971.

Boruch, R. F. On common contentions about randomized field experiments. In G. V. Glass (Ed.), Evaluation Studies Review Annual, Vol. 1. Beverly Hills, CA: Sage Press, 1976.

Campbell, D. T. & Boruch, R. F. Making the case of randomized assignments to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett and A. A. Lumsdaine (Eds.). Evaluation and experiment: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Campbell, D. T. & Erlebacher, A. How regression artifacts in quasi-experimental evaluations in compensatory education tend to underestimate effects. In J. Hellmuth (Ed.) Disadvantaged Child. Compensatory Education: A national debate (Vol. 3). New York: Brunner/Mazel, 1970: 185-225.

Campbell, D. T. & Stanley, J. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

de Charmes, R. Enhancing motivation: Change in the classroom. New York: Irvington, 1976.

Engelmann, S. Basic concept inventory. Follett Educational Corporation, 1969.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-401.

Farr, R. C., Prescott, G. A., Balow, I. H., & Hogan, T. P. Metropolitan achievement tests: Reading instructional tests. New York: The Psychological Corporation, 1978.

Foster, R., Giddan, J. J., Stark, J. Assessment of children's language comprehension. Consulting Psychologists Press, Inc., 1973.

Gall, M. D. The importance of context variables in research on teaching skills. Research in Teacher Education, 1977, 28, 43-48.

Haney, W. A technical history of the national Follow Through evaluation. Sponsored by OE/DHEW, Cambridge, Mass.: Huron Institute, August, 1977.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through. Cambridge, Mass.: Huron Institute, 1974.

Hops, H., Fleischman, D., Guild, J., Paine, S., Walker, H., & Greenwood, C. PEERS: Procedures for establishing effective relationship skills. Unpublished manuscript. Eugene, OR: University of Oregon, 1978.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Monograph No. 1 on Evaluation in Education. Washington, D.C.: U.S. Government Printing Office, 1975.

House, E. R., Glass, G. V., McLean, L. D., Walker, D. E. No simple answer: Critique of the "Follow Through" Evaluation. Harvard Educational Review, 1978.

Kennedy, M. Designing a flexible evaluation: The evaluation of PL 94-142. Paper presented at AERA. April, 1979.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.) Problems in measuring change. Madison: University of Wisconsin Press, 1967.

Miller, J. F. Assessing children's language behavior: A developmental process approach. In R. Schieffelbush (Ed.) Bases of language intervention. Baltimore: University Park Press, 1978.

Newcomer, P. L. & Hamill, D. D. Test of language development. Austin, Texas: Empiric Press, 1977.

Paine, S.& Gersten, R. Direct instruction supervisory code. Unpublished manuscript. Eugene, OR: University of Oregon, 1979.

Peckham, P. D., Glass, G. V., & Hopkins, K. D. The experimental unit in statistical analysis. The Journal of Special Education, 1969, 3, 337-349.

Powers, D. E. & Alderman, D. L. Practical techniques for implementing true experimental designs. Evaluation Quarterly, 1979, 3, 89-96.

Rosenshine, B. & Berliner, D. Academic engaged time. British Journal of Teacher Education, 1978, 4, 3-16.

Stallings, J. A. Follow Through program classroom observation evaluation. Menlo Park, CA: Stanford Research Institute, 1973.

Stallings, J. A. Learning to look: A handbook on classroom observation and teaching models. Belmont, CA: Wadsworth Publishing Company, 1977.

Wolf, R. Review of Metropolitan Achievement Test. In O. K. Buros, (Ed.) The sixth mental measurements yearbook. Highland Park, NJ: The Gryphon Press, 1978.

APPENDIX I

Potential Approaches to Future Research in Follow Through:

Small Scale Studies and Components Analysis

There appears to be a growing consensus that future Follow Through research should contain many small-scale field experiments within the larger evaluation design. The need for small scale, reasonably well controlled studies of manipulation of one or two variables in a field setting was continually expressed at the RFP 101 sponsors' planning meeting in January, 1979. It is a recurrent theme in contemporary evaluation literature (eg., Boruch, 1976; House, Glass, McLean & Walker, 1978; Powers and Alderman, 1979). This is the strategy Kennedy (1979) and her colleagues have adapted in the BEH plan to evaluate implementation of Public Law 94-142.

Small studies offer several important advantages. The smaller scope and sample size could allow for reasonably precise measurement and control of independent variables, such as type of curriculum utilized in the classroom, amount of academic engaged time, extent of parent involvement. In small scale studies, "True" experimental designs - with random assignment of teachers or children to treatment--is a real possibility. Also, various single subject designs - multiple baseline and reversal - are conceivable, using the classroom as the "organism." With either strategy, precise relationships between implementation strategies and teacher/student/parent performance can be investigated. These studies can help isolate the most effective components of successful models, and/or investigate successful components in models which have produced erratic outcomes in different sites. Components from

differing models (eg., Direct Instruction and Parent Education) can be combined to see if even more effective approaches to compensatory education can be developed.

A wide range of studies can be identified to help explore these issues. The research areas discussed in this paper are limited to those compatible with the continuing Follow Through models. This paper will not include discussion of specific research methodologies for the areas of inquiry which are identified. Rather, the focus will be restricted to delineation of research lines which we consider important to the future of compensatory education in general, and of the Follow Through Project in particular.

1. Curriculum and instructional variables. Three proposed studies relate to evaluations of curriculum programs, or to aspects of teacher performance important to the presentation of those programs.

a. Relative effects of two curriculum programs implemented under a common sponsorship. In this study two commercial instructional programs (e.g., SRA's DISTAR Reading and a same-level basal reading program), both implemented under a common model sponsorship (ie., Direct Instruction Follow Through sponsorship), would be compared for their relative effects on relevant process and outcome measures. The study would use the classroom as the unit of analysis, and random assignment of kindergarten classes to reading program. The purpose of the study would be to explore the relative effects of the two academic programs while supervisory support was held constant (i.e., Does curriculum make a difference, or does a good management system alone account for positive results?) Careful fidelity of treatment measures would be essential here.

b. Effects of in-class training procedures on the acquisition and maintenance of critical teacher presentation skills. The goal of this investigation would be to identify and validate effective procedures for training teachers in an in-class situation. The study could be conducted in a manner which would allow for a comparison of the relative effects of training teachers in traditional college classroom or traditional in-service training format versus training them individually, or in small groups (2-3) during the time that they are actually teaching. In this form, the investigation could easily include comparisons of relative costs and time requirements for training, as well as the outcomes achieved by each method. Assessment will include direct measures of specific teacher presentation skills, as well as student outcome measures.

c. Procedures for intervening with low-performing teachers. Some teachers, as well as some students, are low-performers. Like low-performing students, low-performing teachers require intervention if they are to reverse their pattern of failure in school. The achievement of students is too important to allow weak teaching to continue indefinitely. Low-performing teachers might be persons who are capable of teaching effectively but who do not implement an adopted program fully or carefully. They might also be persons who simply have not acquired the skills required for full quality implementation, but who are interested in doing so. Or they might be both unskilled in the model and uninterested in becoming skilled. In any event, research is needed on procedures for bringing low performing teachers to criterion on critical teaching skills. District or program administrators might judge that the low performing teacher is not interested in teaching low income students, and a transfer might be indicated. Or they might conclude that the teacher has potential and

could benefit from training. A number of research questions could be asked regarding training of the low performing teacher. Perhaps the most central of these are what degree of intervention is required to bring low performing teachers to criterion on critical teaching skills; and what procedures produce an acceptable level of teaching and implementation within allowable cost and time constraints? Designs to answer these questions could be similar to those outlined above to deal with more general questions of teacher training. Measure of level of implementation and adequacy in critical teaching skills would provide the data for answering these questions. The work done in the 1978-79 San Diego implementation study, exploring relationships between teacher performance, teacher attitude and student outcome, will offer a good point of departure for these studies.

2. Parents and Families. At least two questions related to parental involvement in Follow Through seem worthy of investigation. These concern parents' ability to (1) impact on students achievement in the classroom and (2) have an effect on the level of program implementation within a site.

a. Effects of direct parental program involvement on student achievement. In this study, parental involvement, in the form of participation in a structured parent education program could be evaluated for its effects on the acquisition of students' academic skills. By implementing such forms of involvement using either a group design, or a multiple baseline fashion across classrooms or schools within a site, the impact on student achievement and social behaviors could be assessed.

b. Effects of models with and without parent education components. Some Follow Through sponsors focus exclusively on the home environment, others exclusively on the classroom. Rather than continuing this either/or dichotomy,

models could be constructed combining elements of both (e.g., a combination of Kansas Behavior Analysis classroom strategies and Chapel Hill Parent Education home visit strategies). This could be contrasted with a classroom-only and a home-only model, preferably all in the same community.

3. Sponsorship Variables. Two studies have been identified which relate to the effects of model sponsorship on program implementation, maintenance, and student performance. These are evaluations of programs conducted with and without sponsorship (in the classical sense), and an investigation of the ways in which FT sponsorship might be gradually reduced over time without producing losses in program fidelity and/or student achievement.

a. Relative effects of program-only and program-plus-sponsorship on implementation process and outcome measures. The purpose of this study would be to evaluate the contribution of model sponsorship to program quality and program effects. This would be done by comparing the implementation of a specific curriculum with no outside sponsorship, and then supplementing that program with active model sponsorship. This comparison could be made using a within-subjects multiple baseline design, with performance at the class or school level serving as a dependent measure. A variety of process and outcome measures would be used to assess the advantages, if any, of adding model sponsorship to the implementation of a standardized curriculum packeage.

b. Identification of procedures and criteria for fading model sponsorship. This study would evaluate the impact on fidelity of treatment and on student performance of gradually and systematically reducing the sponsor's involvement in the implementation of a given model. Implementation would begin with the sponsor taking its usual active role in the sequence. When both process and outcome measures have stabilized under this

arrangement (which is now the case in some Oregon FT sites), fading of sponsorship would begin. Various components of sponsorship - such as frequent site visits, external data management, and on-going technical assistance - would be removed, one at a time, within a multi-element time series design (Sidman, 1960) until only one or two maintenance components remained. The goal of the study would be (a) to identify which components of sponsorship could be faded, and on what schedule, without producing deterioration in either process or outcome measures and (b) to replicate this finding across two or more classrooms, buildings, or sites.

4. Administrative Variables. Several studies are best organized under the heading of administrative variables. These include investigations of the behavior of building principals, local Follow Through directors, other district-level administrators (such as curriculum specialists and superintendents), project managers and sponsor consultants, and Office of Education project officers.

a. Training principals as instructional leaders. This study would provide training for school building principals in instructional leadership and would assess the effects of such training on a) measures of their support for the Follow Through program in their schools, b) indicators of the level of implementation of that program, and c) student performance. These effects likely would best be assessed using a multiple baseline design across schools within a site with possible replication of the procedures across two sites.

b. Effects of management strength by project officers on contract compliance by sites and sponsors. Implementation contracts have been used in Follow Through by some sponsors for a number of years to identify the responsibilities of each party in the program delivery process. However,

such contracts are only as binding as the strength of their potential enforcement. In these studies, data would be collected on strength of project management while measures of contract compliance on the part of both site and sponsor personnel were monitored. Relationships between degree of compliance (or more importantly, in level of implementation) and management strength at the federal level would help to clarify the role of management in controlling program quality. The purpose of this study would be to investigate the effects of strong management by federal project officers in ensuring that both implementing sites and assisting sponsors adhere to the terms of a signed implementation contract.

5. Organizational Variables. Several areas of inquiry can best be categorized as organization or process issues. These include investigations of allocated teaching time and a variety of implementation measures.

a. Systematic variation of engaged time and its effect on student performance. In their review of classroom observational research, Rosenshine & Berliner (1978) found that if students are not adequately taught, they do not learn; i.e., that the amount of academic engaged time is strongly correlated with academic outcomes. A question arising from this finding is, "If they are taught more (or for longer periods of time), do they learn more? And at what level does increasing the amount of engaged time become non-productive?" These questions could be examined by these studies. The amount of academic engaged time for a particular subject area, such as reading, could be systematically increased by varying the amount of instructional time allocated within a multiple baseline design across groups, with replications across classrooms. By measuring student performance on the material being taught (using criterion-referenced test scores in conjunction with standardized achievement test scores), the effects

of varying teaching time could be determined. This area could also be explored via multivariate techniques. The work in the East Las Vegas D.I. site - where Distar is combined with basal readers so that students receive approximately 75 minutes per day of intensive reading instruction could serve as a "pilot study."

b. The relationship between process and outcome data in Follow Through implementations. This work would be an extension of Stallings' (1975, 1977) early work. Multiple regression would be used to analyze the relationship between a set of process variables (amount of academic engaged time, type of corrections, amount of structure, group vs. individual lessons) and a set of criterion measures (standardized achievement scores, some criterion-referenced tests, perhaps language samples).